Analisi dei Generi Letterari della PluriLogia attraverso clusters semantici lessicali

A cura di : Pàncer “Doréamon”, Dæmònya, Adenore Castrovizzardo. In SGRPB7 abbiamo sviluppato del software ed una toolchain per l’analisi quali– e quantitativa dei ***generi letterari***, contenuti in libri, od intere saghe, di <u>genere ibrido</u> o che risultino <u>di difficile catalogazione umana, quantitativa</u>. Un compito arduo al punto che NotebookLM® (Google), lo strumento AI attualmente dotato della più ampia finestra di contesto, ha fallito in maniera invereconda ed esilarante producendo miriadi di allucinatorie … “supercazzole premature” (ci si passi il termine).

Premessa (il SW) :

Il software, di cui condividiamo i sorgenti, sviluppati in linguaggio **GAMBAS 3 per Linux** (pacchetto principale = ***Generatore_Glossari__15_0_0_1.tar.gz*** ed utility integrativa per arrays = ***Integra_Array_0_0_1.tar.gz***) è condiviso con licenza Creative Commons – Attribution : chi vuole (e sa farlo), può modificarlo ed adattarlo alle proprie esigenze, basta che menzioni SGRPB7 e la presente pagina su Archive.ORG.

1) Affinché l’analisi col nostro software abbia successo nel compito dell’analisi semantica dei Generi Letterari occorre che sia stato soddisfatto un prerequisito nei testi da analizzare : <u>la marcatura lessicale dei generi letterari stessi</u>. Un qualsiasi genere letterario di interesse deve essere stato sviluppato nel testo in modo da avere una *fingerprint* non soltanto ad “alto livello”, cioé nei contenuti ed eventi (livello epistemologico), ma anche, <u>**necessariamente**</u>, a “basso livello” (semantico) nella specializzazione nel lessico. Se questa clausola vincolante sia soddisfatta o meno, dipende ovviamente da chi scrive e come scrive ! Nel caso della nostra Saga (***PluriLogia – Una Famiglia di Libri Paralleli***), inclusiva di sei Libri e tre Sequel in corso, il requisito è magnificamente soddisfatto perché essa manifesta in maniera marcata la segregazione lessicale nelle diverse sezioni di testo. È un corpus modulare, stilisticamente un po’ a comparti stagni e che si presta alla *pattern detection*. Lo stile ibrido sperimentale, che internamente chiamiamo lo Stile Disomogeneo, è disomogeneo proprio in virtù della polarizzazione del gergo utilizzato, con iper specializzazione lessicale che investe persino la Voce Narrante stessa. Questo aspetto trascende dalla natura dell’evento narrato nella data circostanza poiché lascia una traccia indelebile nelle singole parole scelte. Il nostro software vede il lessico e le correlazioni tra le singole parti (*token* : vocaboli o singoli frammenti delle parole composte) : dei Libri non vede nient’altro che quello. Se un Autore utilizza un gergo magari amplissimo ma monotono, sempre uguale pur se ricco, un lessico uniforme, generico, per descrivere sezioni molto diverse tra loro nel contenuto sia pure di generi letterari diversi (ad es. un lessico costante in una parte horror, in una romance, in una sezione SciFi oppure fantasy, in una sezione erotica, in una spiritualista / religiosa, in un capitolo weird, in una steampunk piuttosto che dark gothic), allora il SW può percepire solo questo : uniformità, e produrrà grafici piatti, senza clusters, e sostanzialmente inutili.

Una (debole) speranza di cavarne alcunché di buono può ancora rimanere laddove almeno ci siano personaggi molto polarizzati, specializzati in ruoli, menzionati spesso in modo univoco (nome e/o cognome oppure alias esclusivamente propri), sicché si possa mappare la loro distribuzione nel testo. Ma questa mappatura diventa un marcatore di genere soltanto se i personaggi stessi sono stereotipati al punto da risultare “monotematici”. Può succedere, ma è raro, e spesso basta una leggera varianza interna per scatenare falsi positivi nell’analisi e rumore di fondo nel segnale. Quindi, in definitiva, <u>l’esistenza di un lessico specializzato in corrispondenza biunivoca coi generi letterari associati agli eventi, non è opzionale</u>. Ovviamente non occorre che le aree di testo rilevabili siano interamente composte di lessico specializzato. Un 5 % di termini caratterizzanti produce già un buon segnale, ben distinguibile dal rumore di fondo, un 10 % di vocaboli tipici del genere produce un’amplificazione estremamente intensa, per i meccanismi che spiegheremo tra poco.

2) Il secondo requisito è saper <u>identificare in modo adeguato le parole chiave del cluster semantico che costituisce l’ossatura dei Search Terms</u>. Qui il SW non può fare tutto e forse neppure molto (qualcosa sì, però ! Sono presenti un paio di *facilities* utili e di uso intuitivo) : sta all’analista / Autore trovare quali termini lessicali siano associato in modo sistematico e consistente ai vari tipi di generi letterari che sono mescolati nei testi. Se fa male questo passaggio, l’analisi sarà debole od addirittura fallace. Il SW comunque aiuta, con un solidissimo approccio empirico data–driven piuttosto che massimalista e teorico. Ad esempio è totalmente immune al problema, semplicemente monumentale nella lingua italiana molto più che in quella inglese, della LEMMATIZZAZIONE, ossia della derivazione di innumerevoli varianti a partire da un’unica *stem* (radice). Questo perché il SW preliminarmente estrae il dizionario totale usato nei libri, quindi la base dati su cui pascola è l’intero corpus di libri stesso, tokenizzati, atomizzati, e non una stima teorica a priori. Parte da dati reali, *ex post*, e completi. Non ci saranno lemmatizzazioni inutili (previste in teoria ma inesistenti nel testo) né ne verrà persa alcun lemma di quelli effettivamente usati. <u>A patto che la radice sia costante</u>, tutti i suoi lemmi (per nomi ed aggettivi, flessione per genere e numero, per i verbi coniugazione in modi e tempi, inclusa l’aggregazione di particelle pronominali, riflessive etc, e qualsiasi altra specializzazione, inclusi suffissi avverbiali), verranno trovati nella funzione : **CERCA PER RADICE**. Ovviamente, il SW è stupido, non è un linguista : <u>se durante la lemmatizzazione cambia la radice, sta all’autore (come sottolineavamo) inserire tutte le possibili radici di un dato vocabolo pertinente al dato contesto semantico</u>. L’unico limite del SW è che considera **solo lemmatizzazioni terminali**, <u>solo i suffissi. I prefissi no, sono ignorati del tutto</u>. È un limite inaggirabile ma, quantitativamente parlando, fa una differenza così minuscola da essere trascurabile ai fini pratici. Spesso poi ci sono prefissi pericolosi, come l’*alpha* privativo, che sarebbe rischioso equiparare.

Le due fasi dell’Analisi

Prima Fase

La prima fase è l’analisi lineare dei testi. Se si desidera fare analisi incrociate (cross libri), vanno caricati simultaneamente e non in tempi successivi (per una questione di tarature interne). Il solo formato accettato dal SW è puro testo UTF-8. Preparandosi a parte delle righe con un percorso completo di file in ogni riga, è possibile usare il primo dei due (potenti !) comandi BATCH ed analizzare intere Saghe (analisi incrociata concorrente, multilibro). Questa feature ha un significato ontologico molto più profondo di una banale scorciatoia per evitare sequenze ripetitive di comandi ! Ha a che vedere invece con la valutazione globale del parametro più critico (il massimo segnale assoluto dell’insieme, che a sua volta influenzerà le scale relative dello zoom delle ordinate nei grafici) e con la valutazione quantitativa dei clusters di genere letterario, nel contesto di confronti comparativi inter–libri, e non soltanto confronti interni intra–libro. L’analisi quantitativa incrociata è resa possibile dall’utilizzare la medesima scala normalizzata, uniforme sull’intera Saga. Il tempo di elaborazione su circa 3,5 milioni di parole su una macchina del 2020, 6-core model: Intel Core i5-9600K bits: 64 type: MCP cache: L2: 1.5 MiB con 64 GiB di RAM, è dell’ordine dei 5-10 minuti, incluse elaborazioni grafiche (questo dopo averci preso la mano). Non approfondiremo tutto, ma rimaniamo disponibili a fornire assistenza, a tempo perso, per eventuali interessati.

Il punteggio massimo assoluto del Libro ladove il segnale semantico misurato è più intenso viene usato per tarare la scala di tutti i grafici : tutti gli altri Libri saranno ridotti in maniera proporzionale a tale massimo globale posto pari a 100 %. Si vedrà quindi non soltanto la distribuzione locale del dato genere, la diacronia interna, ma la forza comparativa del segnale tra libri differenti. Questa è una vera manna per le Saghe multilibro, rispetto all’analisi singola. Prima però, oltre al dizionario UNIONE (à la Vehnn), che viene estratto automaticamente, bisogna acquistire il Cluster Semantico dei termini di ricerca. <u>Consigliamo fortemente di usare un approccio a stadi successivi, che parta dalle radici e poi un vaglio manuale</u>, piuttosto che la costruzione diretta, interamente manuale, del Cluster da zero.

Inserendo le radici delle parole del Cluster, il SW setaccia il dizionario, raccogliendo tutti i token che cominciano con una delle radici dell’insieme preliminare. Non si usino prefissi brevissimi però, meglio sforzarsi un minimo di gestire le radici capricciose senza delegare tutto al SW, usando radici brevi per sicurezza. In caso contrario, il SW sfornerà clusters preliminari giganteschi, e toccherà revisionarli a mano per sfrondarli dalla presenza di tanti termini estranei al genere che si sta cercando. <u>La revisione manuale resta tassativa quasi sempre, in ogni caso</u>, anche per altre ragioni. Ma farlo su 1000–1500 lemmi è una cosa, farla su 6000–8000 è già diverso.

<u>Questa preanalisi sforna un primo estratto dal dizionario, con le radici già lemmatizzate in ogni modo possibile</u>, ma non possibile secondo le possibilità teoriche della lingua, bensì usate concretamente nel Libro, il ché è equivalente poiché non si manca nessun bersaglio ugualmente e non si sprecano ricerche a vuoto con lemmi inesistenti (che comunque sarebbero da confrontare, rallentando l’esecuzione !). Lavorare in due fasi è più efficiente che pensare di sfornare direttamente il Cluster già lemmatizzato, senza passare per il filtro preliminare della ricerca per radici. Bypassare la funzionalità, da un lato espone al rischio di dimenticare alcuni lemmi arzigogolati (il *si* passivante finale, il *ne o ci* ed altri suffissi contorti, *meli, celo, cela, glieli, gliela*, etc), dall’altro sicuramente per prudenza appesantirà il cluster con voci mai davvero usate in nessuno dei testi, ma che il SW deve cercare allo stesso modo de lemmi esistenti. Si confidi sul fatto che l’analisi multilibro in un colpo crea un dizionario UNIONE : estraendo i lemmi da lì, per radice, non si perde niente.

Intermezzo : refinements ed attribuzione dell’importanza semantica

Successivamente alla creazione del Cluster semantico, serve un altro cruciale sforzo, manuale : <u>catalogare le voci in sub–clusters a differente priorità</u>. I matching delle parole chiave possono, e dovrebbero assolutamente, avere pesi diversi. Bisogna prioritizzare le parole secondo la minore o maggiore pertinenza al genere ricercato. Ci possono essere alcuni termini nel cluster provvisorio ad altissima specificità e tipicità di un dato genere letterario, altri termini piuttosto generici, altri ancora polisemici o dual–use (parole che possono essere usate in contesti specifici, che matchano il dato genere cercato in certi contesti, ma che hanno anche un significato più generale in altri contesti : non bisogna dare a queste eccessiva importanza, perché aumentano i falsi positivi, il rumore di fondo). I primi, più caratterizzanti, dovranno avere pesi superiori, i più generici pesi inferiori. Anche nel classificare i personaggi (i nomi propri sono in sé estremamente specifici ma sono ugualmente parole a rischio a causa dell’alto numero totale di hits dei personaggi protagonisti) bisogna considerare il loro carattere prevalente e la funzione / ruolo nei Libri, se ne hanno. Anche qui : noi di SGRPB7 scriviamo ***Libri di Ruolo*** ed i nostri personaggi sono sovente archetipici, parzialmente o totalmente stereotipati : svolgono funzioni ben delimitate, anche se non tutti e non sempre, ma questo non è vero per tutti gli Autori o per tutti i Libri. Gli altri stiano attenti all’uso dei nomi propri e soprannomi nei clusters semantici : hanno pro e contro. <u>Noi abbiamo incluso spesso alcuni nomi in alcuni generi, ma l’efficacia del SW sulla nostra Saga rispecchia queste asimmetrie di ruoli</u>. Viceversa, una profonda complessità caratteriale dei protagonisti rende consigliabile escluderli dai Clusters, altrimenti si creeranno costellazioni di falsi positivi disseminati ovunque. Melius abundare quam deficiere, in genere, ma non sempre, bisogna conoscere i rischi. Ci sono altri punti di forza nei nomi propri, ma è una cosa da valutare bene ed il SW non può decidere. Un’osservazione : il SW traccia tutto quel che fa in un file di LOG, registra ogni singola *hit*, quale vocabolo ha preso quale punteggio e a che coordinate testuali. Ma è umanamente mostruoso, diversamente dall’analisi dei grafici, gradevolissima e divertente. Diciamo che abbiamo reso così minuzioso il file di LOG perché ci è servito come base per il debug metodologico, durante lo sviluppo. Ora che il SW è a punto e funziona, non lo usiamo praticamente più (salvo per rifinire i cluster semantici, ripescandoli). Anche i dati numerici degli scores e dei counts grezzi vengono salvati come arrays, ma i dati più salienti vengono resi disponibili già nella casella di testo, senza costringere a questo lavoro improbo da minatori.

I Pesi semantici dei sub–clusters.

La gestione dei pesi assegnati ai diversi sub–clusters è assistita dal SW e solleva da calcoli tediosi e da rischi ancor più insidiosi (quello di attribuire una forza arbitraria, a priori, ad un dato genere letterario, diversa da quella attribuita ad altri generi). Ad esempio, nella ricerca del genere EROTICO / PORNOGRAFICO, che è quello che abbiamo scansionato con più minuzia per una ragione molto concreta (stimare i rischi di *detection* da parte di algoritmi filtro “*under age*” e possibili criticità da censura nella pubblicazione su store online), è stato utilizzato un cluster di ben <u>**893 lemmi, suddiviso in 10 sub–clusters con pesi originali 100 90 80 70 60 50 40 35 30 25**</u>. Non abbiate timore nell’usare i pesi : vengono in automatico rinormalizzati internamente in modo proporzionale al valore immesso ed alla sua posizione nel range, e ricondotti ad una scala interna nel range, per così dire, adimensionale **0,1…1,0** in modo che nelle

comparazioni tra segnali di generi diversi, non ci sia un bias di fondo che non avrebbe ragion d'essere. Assegnare pesi grandi o piccoli è irrilevante, tanto vengono riscalati in automatico al range da 0,1 a 1,0 e conta solo la forza relativa. In altri termini le parole del sub–cluster meno specifico verranno poste sempre ad un decimo esatto del peso assegnato al sub–cluster più specifico : ciò normalizza la gamma dinamica totale, la modulazione ed amplificazione del segnale. È abbastanza ragionevole il 10 % di cut off : pesi troppo bassi graficamente non si vedrebbero proprio nelle figure. Come già accennato, se l'analisi è multilibro concorrente, il match delle radici va a pescare dal Dizionario UNIONE dell'intero corpus, e non dall'intersezione comune. Questo implica più lavoro di revisione manuale, ma garantisce di non mancare nessuna hit.

Ampiezza totale del cluster semantico di ricerca

La numerosità assoluta del cluster, ovviamente, influisce sulla forza del segnale rilevato. Ora, se le comparazioni rimangono all'interno di un unico genere letterario, questo non è mai un problema : rimangono possibili sia analisi relative interne (distribuzione nei capitoli o sezioni intra libro) sia esterne (confronti cross libri). Ma se invece interessa confrontare tra loro generi letterari diversi (analisi indipendenti) va tenuto conto che assegnare un cluster vastissimo ad un primo genere A ed uno molto ristretto ad un secondo genere B, introduce un bias di fondo che tenderà a sovrastimare il primo (A) ed a sottostimare il secondo (B). Ciò è inevitabile, a parità di specificità media dei vocaboli, ovviamente ! Tuttavia, al netto della pigrizia del *cluster refinement*, qualora questo capiti, in realtà già di per sé qualcosa racconta del carattere relativo dei libri. Cosa ci dice questo bias "sincero" ? Che se un genere dato si rispecchia in un cluster lessicale enorme, significa che effettivamente quel genere contribuisce tanto o è persino il genere prevalente. Se, al contrario, esso ha un cluster di pochi termini, evidentemente o è stato trattato in modo sciatto, noioso e ripetitivo, oppure dà contributi quantitativamente limitati, molto circoscritti, o persino puntiformi, episodici. Il nostro SW si prefigge di misurare i testi, fornendo dati numerici e grafici quantitativi, non qualitativi. E questo bias sulla numerosità del Cluster di ricerca è già un primo verdetto quantitativo che emerge e da interpretare. Quindi il bias esiste, ma non per un limite del SW, bensì per la natura di fondo dei libri stessi. Non tutti sono ibridi, e non tutti gli ibridi sono misti alla stessa maniera (possono essere molto equilibrati o parecchio asimmetrici : questo si rispecchia nella numerosità dei Clusters di ricerca). Ancora, ribadiamo : melius abundare quam deficiere, ci sono robustissimi argini contro i falsi positivi, li vedremo tra poco. Anche il cluster semantico spiritualista / religioso è stato vasto (**896 lemmi**), quello Felino / Gattoso un po' meno (**306 lemmi netti su 542 preselezionati e 6 soli sub–clusters di pesi relativi 10 9 7 6 5 4**, cfr con 893 di prima !). Il Cluster Vampirico (**151 lemmi in 3 soli sub–clusters**), ci siamo resi conto solo a posteriori, era molto più ristretto non tanto come spazio nomi ma come picchi. Eravamo stati convinti che il genere Vampiresco fosse più pesante nella Saga, ed il software ci sta dicendo il contrario : è presente, significativo, ma relativamente secondario, salvo che nella sola Branca 2. Questo è un verdetto che nemmeno noi, che fummo Autori della Saga (anche del SW però), possiamo ignorare.
Bene : sono stati caricati ed analizzati i testi, scansionati i dizionari ed è stato caricato il cluster, articolato con cura in sub–clusters coi loro debiti pesi, e si lancia l'analisi multi libri. Cosa fa questa procedura ? Lavora in due fasi : distinte in teoria ed algoritmicamente ma percepite in modo monolitico dall'utente che non deve lanciare comandi distinti : è per tale ragione che le accorpiamo nella Seconda Fase, articolandola.

Seconda Fase (l'Analisi vera e propria)

1° step : la raccolta dei dati grezzi (Counts e Scores singoli).

Il SW scorre tutte le parole di tutti i libri (trasformati in sequenze ordinate di tokens), uno alla volta, ed ogni volta che un token (vocabolo del testo) trova un match in uno dei sub–clusters di ricerca, registra tutto. Salva che indice assoluto aveva il vocabolo, in che "blocco" (*chunk di 100 parole : questa scelta è stata obbligata per produrre grafici gestibili, considerato che il libro più lungo, la Branca 1, ha circa 600'000 parole, e produce quindi un'immagine che in ascissa ha oltre 6000 pixel, non proprio piccina*), a che sub–cluster appartiene, e l'indice della parola medesima nel suo sub–cluster. Parallelamente viene salvato in un report (in un file di LOG : il SW lo chiede all'inizio) ogni match (hits) incluso un mini estratto del testo, che parte da "tot" parole prima ed arriva a "tot" parole dopo la *hit*, il pezzetto di libro ricostituito, senza punteggiatura né niente. Questo "tot" è regolabile entro limiti ragionevoli. Se lavorate con Saghe di vari milioni di parole, e fate girare molte procedure in una sessione, il file di LOG lievita presto ai 50 – 100 MB. Può servire per eventualmente raffinare il cluster di ricerca, se si notano troppe *hit* che sono falsi positivi. Approfondire la struttura interna del file di LOG esula da questa mini guida. Per un certo tempo ci eravamo fermati qui, ed i grafici erano abbastanza soddisfacenti ma anche pieni di *rumore bianco*, appunto il disturbo da falsi positivi (le parole più generiche o quelle *dual–use*), che avevano una forza minore ma ancora confrontabile col vero segnale "di genere" da noi cercato.

2° step : attribuzione reciproca, multivoca, dei premi di prossimità.

Più di recente abbiamo introdotto una feature che ci aveva terrorizzato ma che si è curiosamente risolta in 3 – 4 giorni (il resto ha richieso vari mesi di refactoring e debugging), e che ha stravolto la potenza ed efficacia di analisi, introducendo un'amplificazione da sinergie : le analisi dei rapporti di prossimità ed il concetto di finestra mobile (regolabile) di contesto. La nostra Saga ha rivelato una lunghezza media di periodo, inclusi dialoghi ed introspezioni, di 15,75 parole per periodo, arrotondato a 16 parole. Questo è il periodo medio ponderato. Abbiamo definito, arbitrariamente ma con riflessioni statistiche sulla deviazione standard stimata, una context window di 24 parole in ambo le direzioni. Significa effettuare una seconda analisi avanti e indietro, rispetto ad ogni *hit* singola, di 24 parole avanti e indietro (un contesto totale di 48 parole, che copre il triplo del periodo medio, e quindi presumibilmente la stragrande parte anche dei più lunghi : *si noti però che il SW non possiede alcuna nozione sintattica o grammaticale di periodo : per lui l'intero libro è un nastro continuo di parole, nel bene e nel male*). I termini, o *hits*, prima isolati, ora ricevono dei *prizes* incrociati (la cosa è additiva e reciproca, e segue una dinamica combinatoria, totalmente non lineare) se sono vicini ad altri *hits*. Eccovi la funzione di attribuzione dei premi (*prizes*) ad ogni *hit*, che aveva già ricevuto uno *score* base, di partenza, uguale al peso specifico del suo sub–cluster.

```
Public Function FzCalcNormPremioProximity (Distance As Integer, EarlierCoeff As Float, LaterCoeff As Float, Optional SemiWithOfContextWindows As Integer = 24) As Float
   Static conPiGreco As Float = 3.14159265359
   Return (Sqr(EarlierCoeff * LaterCoeff) * 0.5 * (1.0 + Cos(conPiGreco * (Abs(Distance) - 1.0) / CFloat(SemiWithOfContextWindows))))
End
```

Cosa fa ? Riceve dal chiamante la distanza relativa tra una *hit* ed un'altra, purché entro il contesto (in avanti o indietro è indifferente) e la rapporta all'ampiezza totale della *context window* (regolabile). La distanza minima è zero (tra parole adiacenti, non 1) ed il coseno di zero è uno, il valore base massimo del premio. Se la distanza è uguale alla *context window*, l'argomento del coseno tende a PiGreco, ed il coseno crolla a zero. La dinamica non è lineare ma sigmoide : solo nelle zone centrali il valore varia in modo ripido, agli estremi minimo e massimo varia in modo dolce. Questa è la base della premialità di prossimità, che crea una prima **sinergia lessicale** dai termini che si rafforzano reciprocamente, in modo transitivo e biunivoco. Ma non finisce qui : viene anche usata la media geometrica (sqrt (A × B)) dei pesi semantici dei sub–clusters, del termine corrente e del suo vicino (*nella ricerca in avanti, EarlierCoeff è lo hit stesso e LaterCoeff è il peso del vicino in oggetto, nella ricerca all'indietro, EarlierCoeff è il peso del vicino in oggetto e LaterCoeff è lo hit stesso*). La natura dell'interazione lessicale è combinatoria. Se in una data regione un termine è isolato, riceve solo il suo Score base (proporzionale al peso) e stop. Se ha 1 solo vicino, riceve un premio (che restituirà o ha già restituito), perché ci sono 2 combinazioni (il verso conta, fa distinzione). In totale ci saranno 2 premi in quel *chunk*. Se lo *hit* ha due vicini, riceverà due premi e ne darà due. Ci saranno 6 premi distribuiti in totale, internamente al dato chunk (A → B, B → A, A → C, C → A, B → C, C → D). Se ci sono altri 4 vicini (5 parole in totale tutte in prossimità reciproca sufficiente), saranno le C(5,2) a dirci quanti premi totali verranno attribuiti ed accumulati nel chunk : 5! / 2 = 60 premi !! Ovviamente i premi vari saranno diversi tra loro : cambiano le distanze, può cambiare (o può non) la media geometrica dei pesi, ma la dinamica resta non lineare. I *prizes* sono assai flessibili. Se le parole sono generiche, i premi non sono elevati, perché hanno un peso medio basso, ma se sono specifiche, questo smorzamento non c'è o è limitato. La forza dinamica di un segnale denso, fitto, è ben più che additiva, per quello parlammo poc'anzi di sinergia. Questi picchi semantici si auto amplificano rispetto al rumore bianco, si stagliano come guglie sui dintorni se il testo racconta questo, svettano sul rumore di fondo delle parole isolate (anche se singolarmente intense). Sono la *fingerprint* del lessico specialistico di genere che emerge come proprietà di insieme, olistica, vista a medio raggio. Ecco perché la doppia scansione, quella di raccolta lineare di dati grezzi puntiformi e poi la rilevazione delle relazioni di prossimità, ha pulito in modo drammatico i primi grafici pieni di rumore bianco e forte. Sottolineiamo che prima di settare l'ampiezza della *context window* di prossimità, bisognerebbe conoscere il proprio stile e in particolare l'ampiezza del periodo medio espressa in parole. Noi abbiamo scelto 48 parole (24 avanti e 24 indietro) non a caso, ma perché il nostro periodo medio sull'intera Saga era di 16 parole. Se si usano periodi brevissimi, monosillabici, o smisurati alla Tolkien, converrebbe adeguare la *context window* al dato reale e non strafare. Non ha senso considerare prossimi o contigui due termini distanti 20 parole se si scrivono testi lirici con strofe di poche parole.

Curiosamente avevamo tentato di usare i filtri audio di Audacity (infatti ci sono funzioni di sintesi di file .WAV ed estrazione di campioni nell'utility integrativa per arrays = ***Integra_Array_0_0_1.tar.gz****), che avevano fallito catastroficamente e quella filiera è stata abbandonata in favore di un approccio più "epistemologico". La FFT (fast Fourier transform) distingue bene segnale e rumore se sono spettralmente ben separati : qui non lo sono affatto, il rumore bianco ha frequenze solo di poco superiori ad eventuali picchi dei clusters veri, e nei patterns a macro Cluster multipicco, addirittura le frequenze sono identiche. Inoltre anche il rapporto di intensità relativa non era abbastanza favorevole per una pulizia di stile audio (il volume dei click e del fruscio deve avere un guadagno sensibilmente inferiore al segnale, o si spiana tutto). L'occhio vedeva qualcosa, la matematica non ancora, il modello puntiforme non coglieva il messaggio nel testo. L'approccio non sonoro, ma semantico, ha asfaltato tutto il rumore come per magia, senza perdere nessuno dei picchi intravisti prima, anzi amplificandoli in modi efficacissimi. Si vede proprio il testo, zone di aumento, di introduzione, e di uscita sfumata.*

Risultati delle Analisi comparative

L'analisi semantica supportata dalla rilevazione delle relazioni di prossimità entro la *context window* ha fatto emergere disomogeneità interne a certi libri ed esterne (inter libri) ben più marcate di quanto già non sospettassimo, e la cosa ci ha anche fatto un enorme piacere, proprio perché l'ibridazione di genere sperimentale e la non omogeneità stilistica era una delle nostre scelte progettuali, un vero e proprio 'manifesto' della composizione a più mani. Non abbiamo interrogato tutti i generi dichiarati. I due generi, di gran lunga più intensi e caratterizzanti, il genere **FANTASY** e quello **FANTASCIENTIFICO** (**SteamPunk "sovietica"**), non sono stati mai nemmeno presi in considerazione. Essi sono praticamente ovunque ed a volume altissimo, uniformemente altissimo, non sarebbe emerso granché. Altri generi (es. HORROR) sono stati trascurati in quanto ritenuti a priori (a torto o a ragione) marginali. Sono stati analizzati solo quattro generi letterari complementari : **EROTICO** / PORNOGRAFICO, **GATTOSO** / FELINO, **RELIGIOSO** / SPIRITUALE / MISTICO / ULTRATERRENO / METAFISICO, **VAMPIRICO**. E purtroppo, con enorme rammarico, è stato trascurato per incapacità o impossibilità, uno che avremmo desiderato poter analizzare : il **WEIRD** (STRANO / ECCENTRICO / BIZZARRO / MISTERO, con tracce "*GOTICHE*").

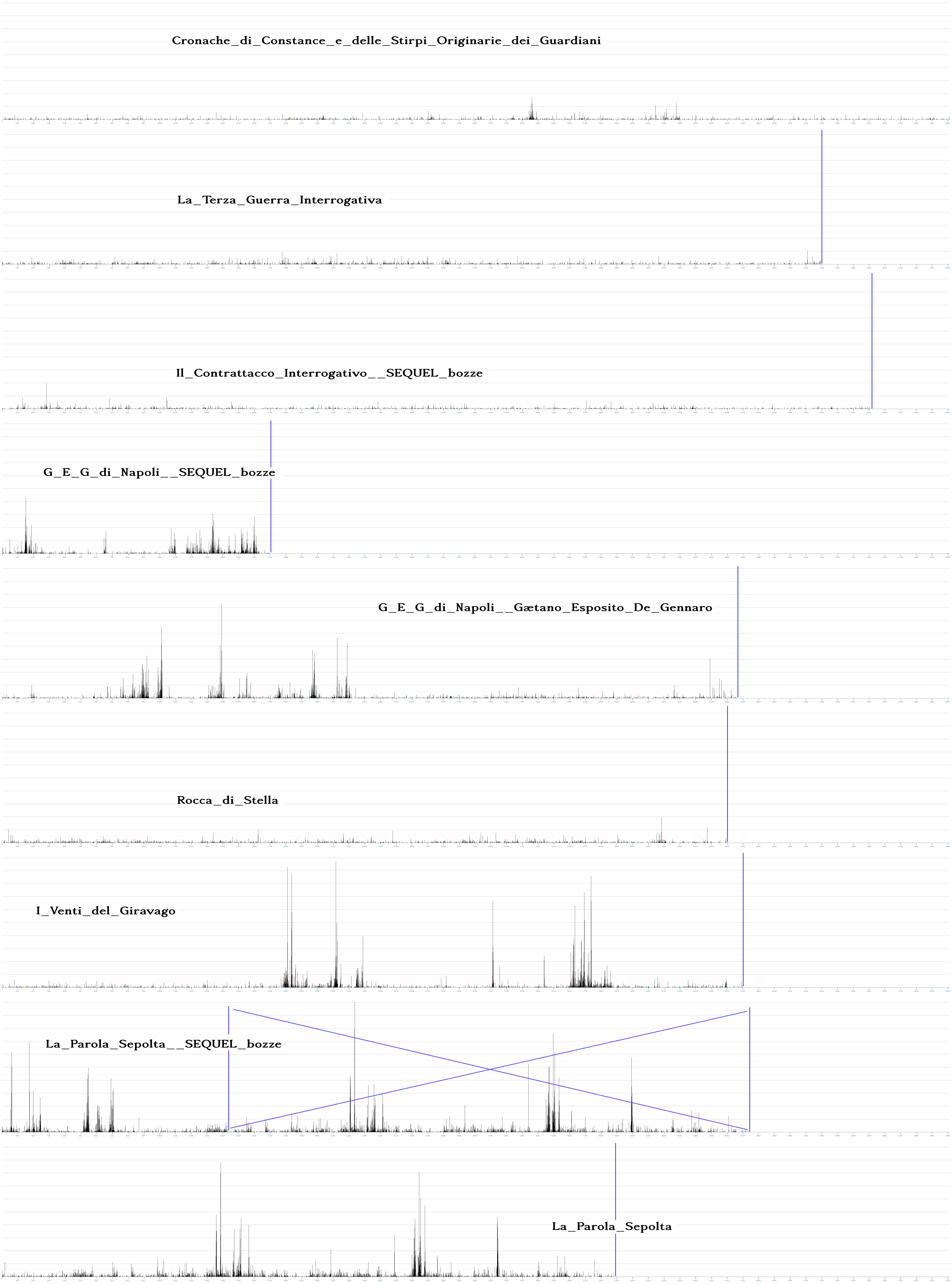
IL GENERE EROTICO / PORNOGRAFICO

Forte del più vasto tra tutti i clusters semantici di ricerca utilizzati, inclusivo di alcuni termini *dual–use* estremamente generici (è stato deciso di massimizzare la raccolta a scapito della selettività, accettando quindi un elevato numero di falsi positivi), aggettivi o sostantivi come : *rossa / rosse, sinistra / sinistro, miss, ragazza, signorina* (che più aspecifici e polivalenti non si può), e persino alcuni nomi propri : Shænnon Hoelmstædt (adeguato, è una comparsa), Sàmaranth (meno adeguato : è menzionata molto più spesso in contesto militare e geopolitico che non in quello erotico, pur essendo l'unica figura erotica della sua Branca), Lesna Balögh – Claire Marie Eloise Skelloer – Llyrahl Ahmoogh (poco adeguati : sono menzionate molte volte, anzi in prevalenza, al di fuori del contesto misurato, e tante volte anche in senso assoluto, gonfiando di certo i falsi positivi di un setaccio a strascico siffatto), Àlorie Alorie Jalix Wholle

(assolutamente marginale in questo contesto, nondimeno raccolta ugualmente), ha prodotto un segnale intenso. Con un simile cluster semantico la nostra prudenza ha raggiunto il livello paranoide per lo scopo originario : emulare l'*automatic detection* algoritmica di Amazon, You Tube etc sui contenuti "*under age*" 18+. Abbiamo quindi scientemente accettato l'amplificazione anche del rumore di fondo e la sovrastima del segnale come più garantista di una sottostima. Ed in effetti il volume del segnale, così amplificato, è salito, non tanto sui picchi massimi, assolutamente veritieri degli episodi hot, quanto sul dato della media, eccessivo.

Legenda delle Branche, incluse Bozze e SEQUEL	Dati del Cluster EROTICO / PORNOGRAFICO
1) Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani 2) La_Terza_Guerra_Interrogativa 3) *Il_Contrattacco_Interrogativo* 4) *SEQUEL__G_E_G_di_Napoli* 5) G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro 6) Rocca_di_Stella 7) I_Venti_del_Giravago 8) *SEQUEL__La_Parola_Sepolta* 9) La_Parola_Sepolta	Riga 1 : 6014 elem., Score Tot. : 1165,9828; Max. di Riga = 11,9182 – media : 1165,9828 / 6014 = 0,19387808447 Riga 2 : 5200 elem., Score Tot. : 1274,2264; Max. di Riga = 6,4224 – media : 1274,2264 / 5200 = 0,245043538462 *Riga 3 : 5516 elem., Score Tot. : 936,6417; Max. di Riga = 11,3797 – media : 936,6417 / 5516 = 0,169804514141* *Riga 4 : 1694 elem., Score Tot. : 1394,8124; Max. di Riga = 26,1964 – media : 1394,8124 / 1694 = 0,82338394333* Riga 5 : 4658 elem., Score Tot. : **2476,2113**; Max. di Riga = 45,4248 **– media : 2476,2113 */ 4658 = 0,53160397166*** Riga 6 : 4608 elem., Score Tot. : 1252,3787; Max. di Riga = 11,6701 – media : *1252,3787 / 4608 = 0,27178357205* Riga 7 : 4701 elem., Score Tot. : **2950,2025**; Max. di Riga = 65,9022 **– media : *2950,2025 / 4701 = 0,62756913423*** *Riga 8 : 4733 elem., Score Tot. : 4583,3418; Max. di Riga = 65,744 – media : 4583,3418 / 4733 = 0,96837984365* Riga 9 : 3882 elem., Score Tot. : **3320,8739**; Max. di Riga = 57,4349 **– media : *3320,8739 / 3882 = 0,85545437919*** Totali : 41006 elem., **Score Tot. Saga + SEQUEL : 19283,1132 – media :** *19283,1132 / 41006 =* ***0,470251016924*** Massimo assoluto : **65,9022** Range relativo delle medie : (*0,96837984365 – 0,169804514141) / 0,470251016924 =* ***169,82 %***

Ogni elemento (elem.) equivale in tutto il seguito ad un *chunk* di testo di 100 parole, e nei grafici privi di zoom, a dimensione originale (non quindi i presenti, scalati), ad un pixel orizzontale. Le righe in blu rappresentano la fine del testo, perché le figure sono tutte larghe 6014 campioni (la Branca 1, la più lunga), ma tutte le altre sono più corte. Le parti piatte a destra di ogni riga blu non sono parti di testo vuote, a zero *hits*, ma significano che il testo era terminato prima.



Esegesi grafica → testuale. NdR : nelle scale in *chunks* in blu, ogni tacca del righello sono 10 unità, e siccome ogni unità sono 100 parole esatte (senza interpunzioni), allora una tacca sono 1000 parole. 1000 parole, equivalgono a circa da un terzo ad un decimo di capitolo medio o 62,5 paragrafi medi, nella PluriLogia.

1) **Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani** : Il genere EROTICO / PORNOGRAFICO è estremamente debole, presente in 2 – 3 micro clusters, di durata variabile ma intensità assai bassa, nei chunks 3340 – 3390 (il + intenso : protagonista Principessa Sàmaranth), poi uno debolissimo a 3615 – 3725 (stessa protagonista) ed infine un più prolungato a 4070 – 4300 (sempre Sàmaranth : unico personaggio con carica erotica dell’intero Libro). NdT

2) **La_Terza_Guerra_Interrogativa** : Il genere EROTICO / PORNOGRAFICO è praticamente ed uniformemente assente dal libro. Tracce impercettibili di un micro cluster piuttosto lungo (protagonista Lady Janice la spia delle Ali Nere

3) **Il_Contrattacco_Interrogativo** : ancora in stato di bozza ma già di volume colossale. Il genere EROTICO / PORNOGRAFICO è praticamente ed uniformemente assente dal libro (come nel PREQUEL), salvo un breve e molto debole micro cluster iniziale (con protagonista Àlorie Jalix McWholle

4) **SEQUEL__G_E_G_di_Napoli** : nessun commento, sono solo BOZZE analizzate in anteprima

5) **G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro** : Genere EROTICO / PORNOGRAFICO intenso e piuttosto diffuso (copre tra 1/3 e 40 % del libro), sebbene in prevalenza distribuito nella prima metà scarsa del libro, con anche una intro del 12,5 % con un singolo brevissimo episodio (sorellastra di G.E.G., Ambra). Dal 13 % al 48 % del libro si sussegue una serie di clusters talvolta raggruppati in macro clusters. Un crescendo di 6 / 7 clusters in pattern di macro cluster, con baricentri delle bande a 580 – 680 – 760 – 830 – 85 – 1000 – (1060), poi un altro cluster singolo, il più intenso del libro da 1310 a 1410, poi uno più debole ma prolungato da 1500 a 1580. uno basso ma duraturo e "denso" a 1730 – 1820, uno densissimo a 1950 – 2020, uno doppio a 2095 – 2240 ed uno finale, debole e rarefatto a 4490 – 4640. Tutti gli episodi hanno come protagonisti G.E.G. medesimo e la Duchessa, Lady Claire Marie Eloise Skelloer. Globalmente raggiunge tra il doppio e quasi il triplo dei 4 Libri più "deboli" ed anche il suo picco massimo è il secondo classificato, il primo posto venendo assegnato ex æquo a due libri con record distinti. Globalmente il genere erotico è rappresentato nel 44,3 % della superficie testuale.

6) **Rocca_di_Stella** : Genere EROTICO / PORNOGRAFICO molto debole ed estremamente circoscritto : un unico micro cluster tra il chunk 4110 e 4210 (basso), un singolo evento puntiforme a 4475 (protagonisti Kaern Yoght e Lady Shænnon Hoelmstædt). È sostanzialmente equivalente a **La_Terza_Guerra_Interrogativa**, ed al primo (salvo che questo è più diluito a causa della mole titanica)

7) **I_Venti_del_Giravago** : Genere EROTICO / PORNOGRAFICO molto intenso e diffuso, anche se circoscritto alla regione mediana (inizio) : la frazione totale di testo in cui è presente il genere è del 45 % (inizia a 37,65 % e si chiude a 82,54 %), bizzarramente una frazione praticamente identica a G.E.G. ! È il secondo classificato per somma cumulativa ed il primo per picco più intenso. Pattern macro cluster vasto e multi picco (5) tra 1770 e 2300 (include i picchi principali e record assoluti di intensità dell'intera Saga, esclusi i SEQUELA). Poi un cluster debole a 2785 – 2815, alcuni isolati sparsi : 3117 (forte), 3155, 3210, 3440 medi, un ultimo gigante, vasto, denso ed alto quasi quanto il record, tra 3605 e 3880. Protagonisti, praticamente esclusivi Varel Kajinark e Miss Lesna Balögh la Vnýdzyr. È on par col seguente il più rovente della PluriLogia.

8) **SEQUEL__La_Parola_Sepolta** : nessun commento, sono solo BOZZE analizzate in anteprima :

9) **La_Parola_Sepolta** : Genere EROTICO / PORNOGRAFICO molto intenso e diffuso, non detiene il record di intensità assoluta (è secondo), ma detiene quello per score totale (è primo in punteggio cumulativo), perché i suoi clusters sono tra i più fitti e vasti, e ci sono anche micro cluster (bassi, ma ancora densi e larghi, che non appaiono come falsi positivi proprio a causa della duplice signature di larghezza e fittezza : questo sottobosco cova nella parte iniziale, in quella media e in quella finale. È piuttosto diverso da 5) e 7) in questa maggiore pervasività e minore segregazione in episodi polarizzati. Amplissime regioni di sottobosco, basso a fitto : 310 – 580, 620 – 730, 820 – 890, 910 – 1060, 1075 – 1260, 1690 – 1815, poi ancora sottobosco 2990 – 3880 con incluso un superpicco densissimo e stretto a 3140 – 3150. Un macro cluster con pattern a multi picco, vasto e intenso, a 1290 – 1670, ulteriore sottobosco tra 1950 e 2500 con immersi 4 picchi medi o forti : 1990, 2085, 2325 e 2490. Il suo primato di score totale è ancora più rimarchevole in quanto è anche il libro più corto (se si escludono le bozze), quindi l'intensità media di questi cespugli non va sottovalutata come magari l'occhio potrebbe fare sul grafico : i numeri dicono altro.

Il GENERE GATTOSO / FELINO

Come già menzionato, il Cluster semantico Gattoso / Felino, ha un bias di fondo, rispetto a quello erotico, di uno spazio complessivo di lemmatizzazione di soli **306** termini netti contro **893** del precedente. Questo in teoria rende di per sé molto meno efficiente la pesca a strascico basata su un filtro ampio di selezione. Una rete grande appena un terzo, inevitabilmente raccoglie meno pesci. I lemmi del leitmotif felino sono inoltre mediamente più specifici, infarciti come sono di neologismi con frequenze assolute pari a zero al di fuori delle aree tematiche di genere. Questi neologismi, essendo licenze poetiche e sgrammaticature, tendono a non essere usate con generosità nel testo neppure da una Voce Narrante assai "schierata". Sono assenti nelle sezioni non specializzate, ed usate con parsimonia anche in quelle di genere felino. Questo significa un ridotto segnale spurio, un minor numero di falsi positivi. Un cluster felino è significativo anche se basso, purché sia almeno fitto oppure esteso. Ora lasciamo parlare gli outcomes (i grafici comparativi), per confrontarli col genere precedente.

<table>
<tr><td>Legenda delle Branche, incluse Bozze e SEQUEL</td><td>Cluster GATTOSO / FELINO</td></tr>
<tr><td>1) Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani
2) La_Terza_Guerra_Interrogativa
3) Il_Contrattacco_Interrogativo
4) SEQUEL__G_E_G_di_Napoli
5) G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro
6) Rocca_di_Stella
7) I_Venti_del_Giravago
8) SEQUEL__La_Parola_Sepolta
9) La_Parola_Sepolta</td><td>Riga 1 : 6014 Elem., Score Totale : 1883,5181; Max. Riga = 22,4493 - media : 1883,5181 / 6014 = 0,313188909212
Riga 2 : 5200 Elem., Score Totale : 1282,4735; Max. Riga = 46,4691 - media : 1282,4735 / 5200 = 0,246629519231
Riga 3 : 5516 Elem., Score Totale : 4729,2214; Max. Riga = 53,257 - media : 4729,2214 / 5516 = 0,857364285714
Riga 4 : 1694 Elem., Score Totale :   814,2409; Max. Riga = 32,1431 - media : 814,2409 / 1694 = 0,480661688312
Riga 5 : 4658 Elem., Score Totale : 1737,4837; Max. Riga = 22,0138 - media : 1737,4837 / 4658 = 0,373010669815
Riga 6 : 4608   Elem., Score Totale : 1965,8584; Max. Riga = 52,2769 - media : 1965,8584 / 4608 = 0,426618576389
Riga 7 : 4701   Elem., Score Totale : 2430,2949; Max. Riga = 34,8507 - media : 2430,2949 / 4701 = 0,5169740268
Riga 8 : 4733   Elem., Score Totale : 1060,8701; Max. Riga = 38,3866 - media : 1060,8701 / 4733 = 0,224143270653
Riga 9 : 3882   Elem., Score Totale :   861,2530; Max. Riga = 38,3866 - media : 861,253 / 3882 = 0,221858062854

Totali : 41006 elementi. Somma Totale : 16765,214 - media : 16765,214 / 41006 = 0,408847827147
Massimo assoluto : 53,257
Range relativo delle medie : (0,857364285714 – 0,221858062854) / 0,408847827147 = 155,44 %</td></tr>
<tr><td colspan="2">Comparazione immediata col cluster EROTICO / PORNOGRAFICO : (Score Tot. Saga + SEQUEL : 19283,1132 – media : 19283,1132 / 41006 = 0,470251016924 – Massimo assoluto : 65,9022). Si rammenti che il "setaccio" del cluster gattoso / felino è vasto appena un terzo di questo, e ciononostante raccoglie score del tutto confrontabili (talvolta superiori).
Il genere GATTOSO / FELINO quindi è moderatamente (significativamente ?) più uniforme del precedente.
Cfr. con : Range relativo delle medie : (0,96837984365 – 0,169804514141) / 0,470251016924 = 169,82 %</td></tr>
</table>

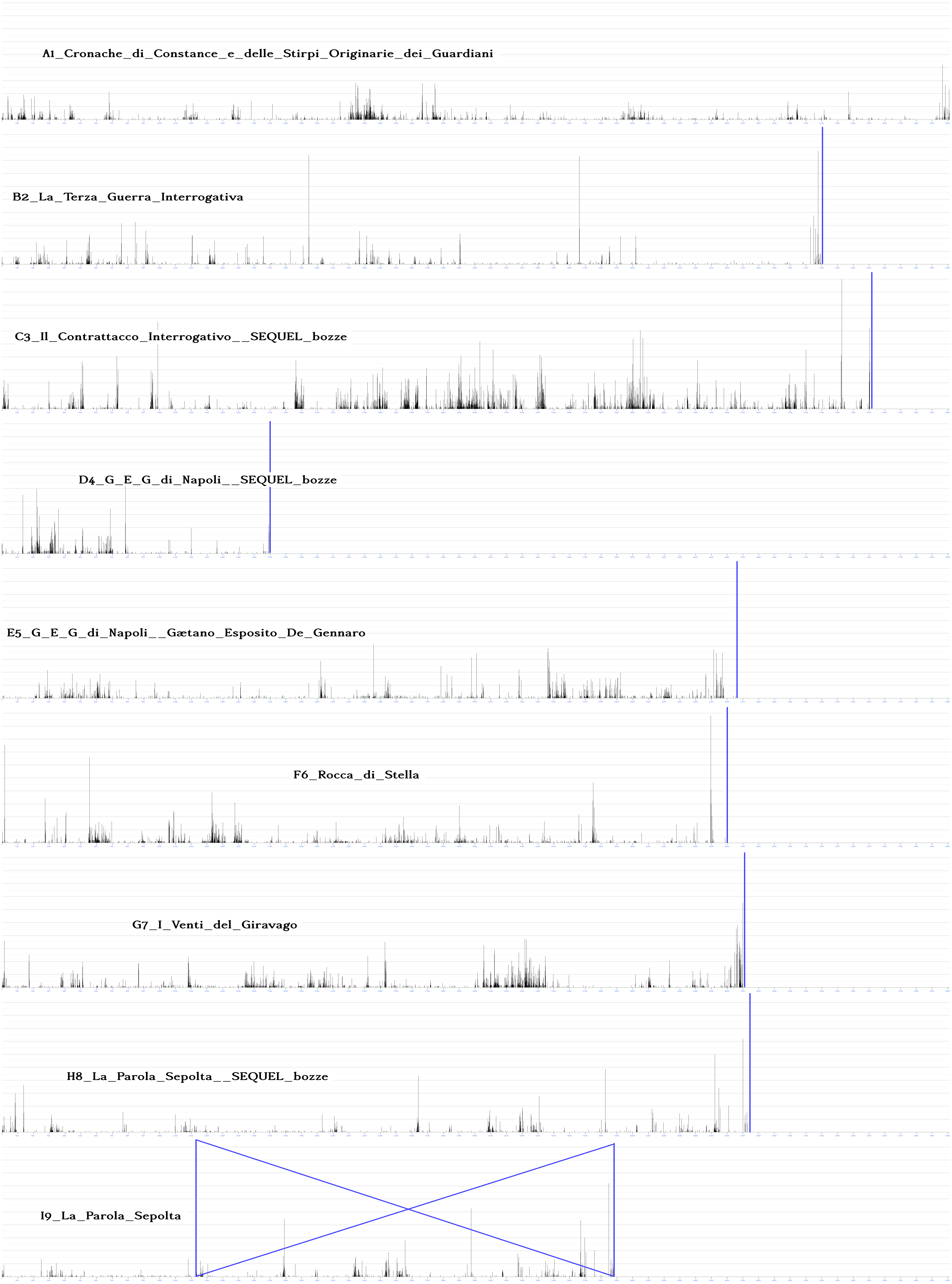
Nota importante : l'inclusione delle Bozze del *SEQUEL__La_Parola_Sepolta* ha peraltro alterato abbastanza le medie del cluster EROTICO (aumentandole !), per il semplice fatto che i due terzi di quel testo contengono ancora una copia intera del Prequel per agevolare la consultazione (ovviamente saranno rimosse !), e questo ha comportato di includere due volte il Prequel stesso nella statistica, sovrastimandone il peso (visto che era già il Libro col segnale EROTICO più intenso !). Le Bozze nuove in sé sono appena un terzo dei 3882 *chunks* (in realtà la figura è persino errata : un cluster molto piccolo del prequel non è stato riconsociuto : le Bozze nuove sono ancora meno). Il punto rimane pertanto che, al netto di ciò, a fronte di un cluster semantico molto più ristretto e meno afflitto dai falsi positivi per la minore incidenza di parole generiche e *dual–use* (le quali non non mancano *tout court*, ma sono oggettivamente meno numerose e tutte collocate negli ultimi due sub–clusters a peso inferiore, i quattro alti sono estremamente specifici), il genere GATTOSO / FELINO fa a pugni nella stessa categoria del precedente. In altri termini, al di là della distribuzione diversissima (assenza dei colossali patterns a macro cluster multipicco rilevati sopra), il genere GATTOSO / FELINO compete *on par* con il rivale che aveva beneficiato di una rete a strascico iper prudente, paranoide, vasta tre volte tanto. Il GATTOSO gioca nel sottobosco, insinuandosi come la gramigna un po' dappertutto, tenacemente. L'intera serie dei 9 grafici riportati è stata schiacciata dai due massimi isolati sopra score 50 di Rocca_di_Stella e Il_Contrattacco_Interrogativo (La_Terza_Guerra_Interrogativa è un pochino inferiore). Questi massimi sono isolati ed hanno l'effetto di demoltiplicare tutti gli altri.

I clusters fitti / vasti nella Saga. In Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani se ne contano 3 – 5 sparsi un po' per l'intero testo, anche se in prevalenza nella prima metà. La_Terza_Guerra_Interrogativa ha un pattern macro peculiare : niente "sottobosco", tanti picchi isolati (anche alti) distribuiti per quasi l'intero testo salvo la zona di Space Opera del finale (ed un ultimo altissimo picco proprio in chiusura del Libro). Il_Contrattacco_Interrogativo giganteggia su tutta la Saga e potrebbe quasi intitolarsi il libro dei Miagoliéri. È un libro assolutamente polarizzato dove questo genere dilaga per tutta la trama, inanellando sia il record di score totale di tutti i generi sia il picco di genere più alto della Saga.

G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro come estremali è molto simile a Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani, salvo la presenza di una più estesa regione "vuota" alla fine della prima metà del Libro.

Rocca_di_Stella ed I_Venti_del_Giravago, entrambi superiori a GEG, mostrano verdetto misto. Rocca_di_Stella presenta un picco massimo di altezza decisamente superiore, ma I_Venti_del_Giravago ha uno score totale cumulativo nettamente più elevato (è, in effetti, il secondo dell'intera Saga. Questo ci ha sorpreso : avendo visto quanto fosse marcata la signature del genere EROTICO, dove arrivava secondo nella Saga, non ci aspettavamo che arrivasse secondo anche qui !). Bizzarramente Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani, che è in assoluto il libro più lungo e di molto, è piuttosto debole in entrambi i generi. Vedremo se avrà la sua rivincita nel genere RELIGIOSO (il genere hard SciFi, laddove giganteggerebbe, come già detto non verrà nemmeno analizzato).

Ne La_Parola_Sepolta il genere felino è debole, presentando clusters meno estesi, meno densi internamente, più rarefatti e con picchi massimi al meglio di media intensità. In effetti, tranne le 4 Forfére fantasma nell'Æthéreum, il tema GATTOSO è secondario in questo Libro. Anche qui, nell'immaginario dei più di noi, questo Libro era simile al precedente. I numeri non sembrano pensarla così. E i numeri sono oggettivi, non opinioni. Il bello di questo SW è assistere nell'auto consapevolezza del testo.



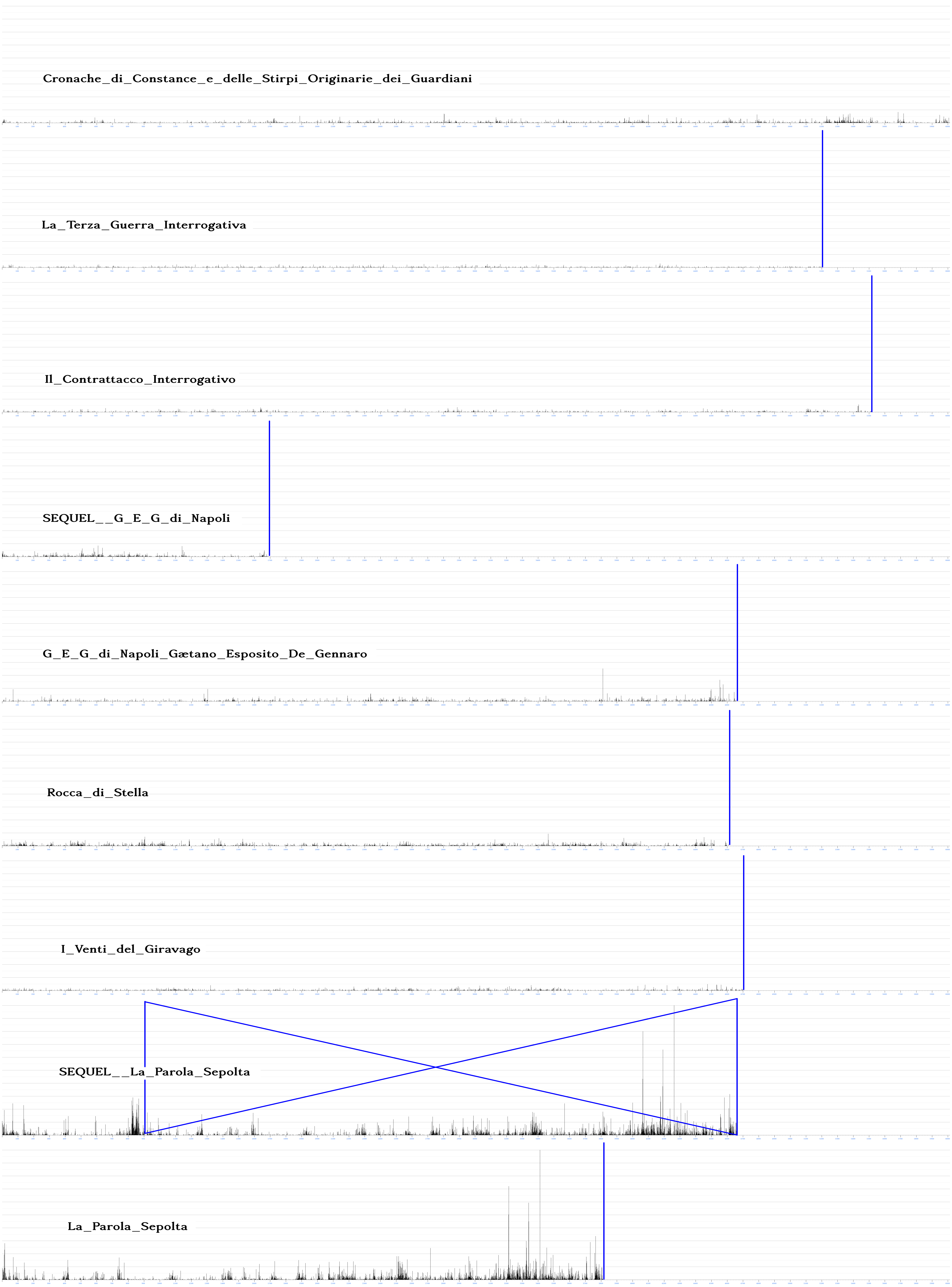
IL GENERE RELIGIOSO / SPIRITUALE / MISTICO / ULTRATERRENO / METAFISICO

Conviene sgombrare il campo da alcune peculiarità nella misura di questo genere, più esattamente sul cluster semantico servito per la raccolta del segnale testuale. **Primo**, sono state praticamente espunte le cosiddette religioni apocrife ed il loro gergo specifico se presente e discernibile. Questo è successo col culto del Sire Ardente / Sire Fiameggiante (anche detto il Titano al Centro del Mondo) e le parole correlate al suo culto ed alla Piromanzia. Nessuna di queste keyword è entrata nel cluster di ricerca, penalizzando fortemente Rocca_di_Stella, ma ciò è stato intenzionale : questa entità non era un vero Dio, bensì un Alieno, seppure incommensurabile e quindi adorato, essendo percepito come un Dio dai primitivi Hochàim. Si è data invece enfasi alle *nostre* religioni terrestri (in particolare Abramitiche), salvo laddove il gergo specifico fosse condiviso. **Secondo**, da La_Parola_Sepolta e dal suo SEQUEL, è stato necessario rimuovere i glossari, in cui la densità di termini e l'algoritmo usato (sinergia di prossimità) faceva esplodere i punteggi asfaltando il segnale di tutti gli altri libri. Ora è vero che questa Branca rimane in ogni caso la più intensa da questo punto di vista e anche senza sovrasta gli altri Libri, tuttavia l'ordine di grandezza della sua superiorità e la distribuzione dei massimi aveva destato sospetti.

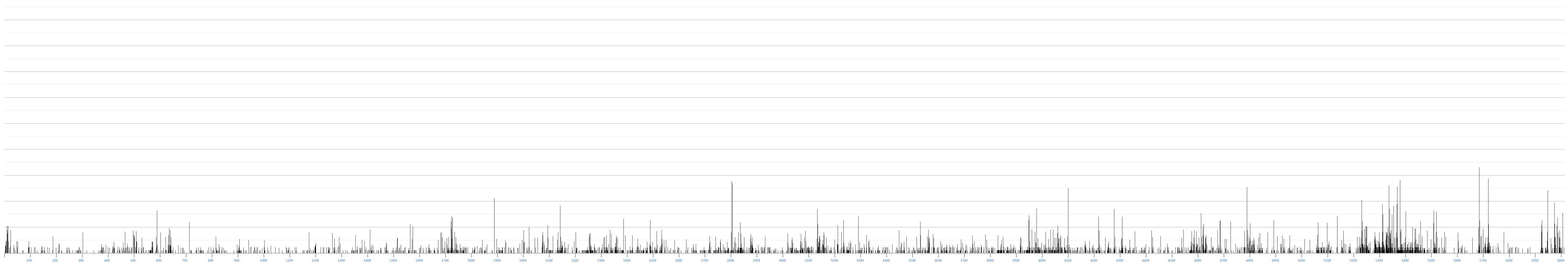
Un riesame, specialmente grafico, aveva mostrato un colossale picco nelle ultime decine di pagine di entrambi i libri, e nasceva semplicemente dall'Appendice Demonologica presente nel Glossario (presente in entrambi), che conteneva concentratissimi tutti i vocaboli più specifici del cluster, sicché questa breve sezione che mandava in tilt l'algoritmo è stata scientemente espunta dal testo. Nonostante questa precauzione, i due libri del filone de La_Parola_Sepolta continuano ad essere quasi fuori scala, confermando la fortissima matrice RELIGIOSA presente.
Oltre al catastrofico massimo locale (159 ! essendo il secondo classificato appena a 40 ed il terzo a poco meno di 15, ossia un decimo scarso del primo, che segnala l'asimmetria davvero imponente), ciò che più si rivela diagnostica è la media pesata per numero di text *chunks* (relativamente pochi qui : il libro più breve è appena il 63,735 % di Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani, e ciononostante giganteggia). Il segnale di Libro è 3 volte la media e ben 12,3 volte il Libro più debole (La_Terza_Guerra_Interrogativa).
Il cluster semantico usato nelle misure è molto vasto (con ben **897 lemmi, suddiviso in 7 sub-clusters di pesi : 100 90 80 65 50 40 35**) ed al contempo mediamente ancora più specifico degli altri (cosa che riduce il rumore di fondo).

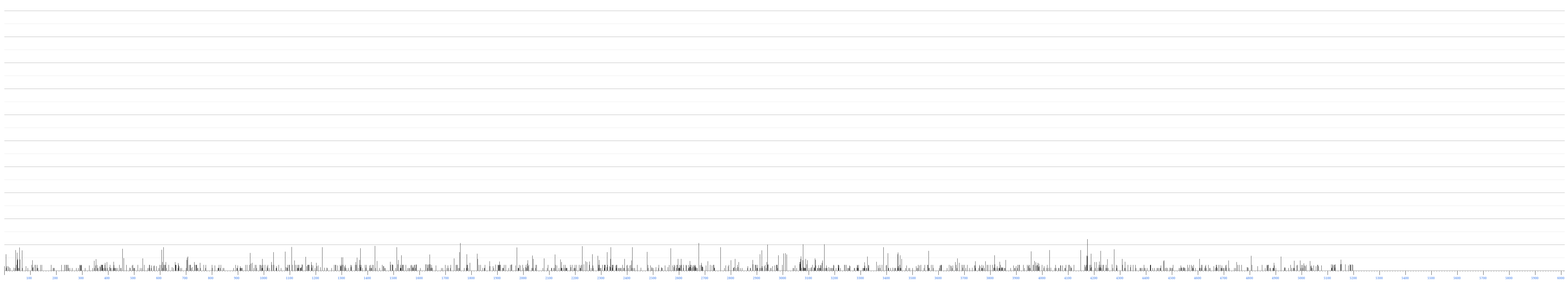
Legenda delle Branche, incluse Bozze e SEQUEL	Cluster RELIGIOSO / SPIRITUALE / MISTICO / ULTRATERRENO / METAFISICO
1) Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani	6014 elem. Somma Parziale : 2467,3687; Max di Riga = 13,2868 - media : 2467,3687 / 6014 = 0,410270818091
2) La_Terza_Guerra_Interrogativa	5200 elem. Somma Parziale : 967,8034; Max di Riga = 4,8546 - media : 967,8034 / 5200 = 0,186116038462
3) *Il_Contrattacco_Interrogativo*	5516 elem. Somma Parziale : 1227,9801; Max di Riga = 9,3154 - media : 1227,9801 / 5516 = 0,222621482959
4) *SEQUEL__G_E_G_di_Napoli*	1694 elem. Somma Parziale : 951,3636; Max di Riga = 13,5217 - media : 951,3636 / 1694 = 0,561607792208
5) G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro	4658 elem. Somma Parziale : 2053,371; Max di Riga = 40,103 - media : 2053,371 / 4658 = 0,440826749678
6) Rocca_di_Stella	4608 elem. Somma Parziale : 2927,0365; Max di Riga = 14,9314 - media : 2927,0365 / 4608 = 0,635207573785
7) I_Venti_del_Giravago	4701 elem. Somma Parziale : 1493,4746; Max di Riga = 7,8703 - media : 1493,4746 / 4701 = 0,317692958945
8) *SEQUEL__La_Parola_Sepolta*	4683 elem. Somma Parziale : 10484,1849; Max di Riga = 159,0133 - media : 10484,1849 / 4683 = 2,23877533632
9) La_Parola_Sepolta	3833 elem. Somma Parziale : 8760,3615; Max di Riga = 159,0133 - media : 8760,3615 / 3833 = 2,28551043569
	Totali : 40907 elementi. **Somma Totale : 31332,9443 - media** SAGA : 31332,9443 / 40907 = **0,765955565062**
	Max assoluto : 159,0133
	Range relativo delle medie : (2,28551043569 – *0,186116038462) / 0,765955565062 =* ***274,1 %.***

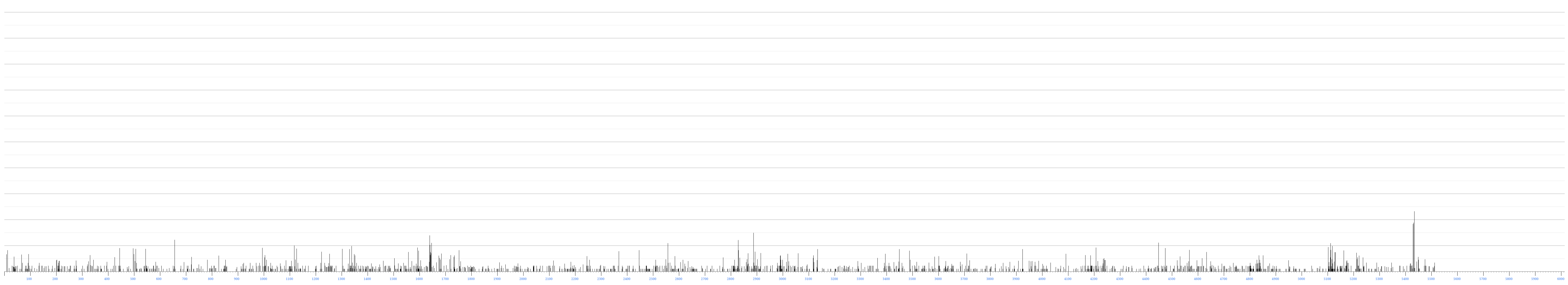
Di gran lunga il più eterogeneamente distribuito tra i generi rilevati !

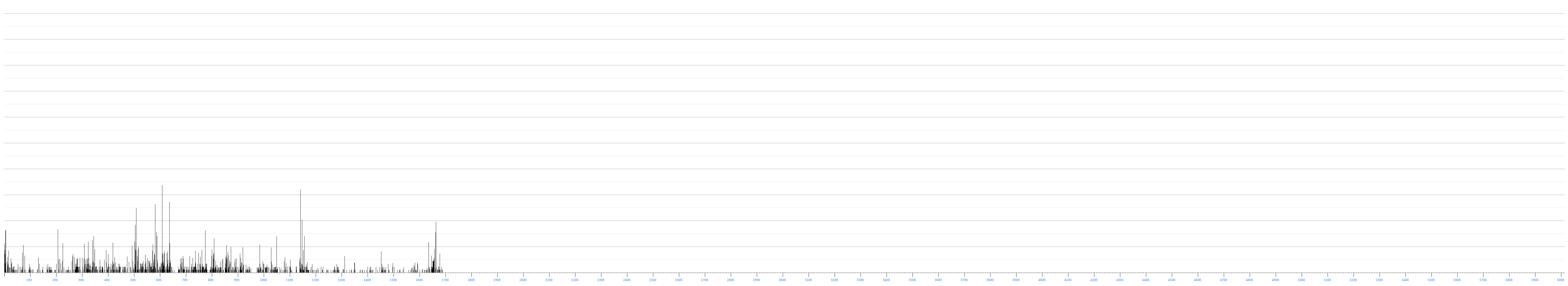


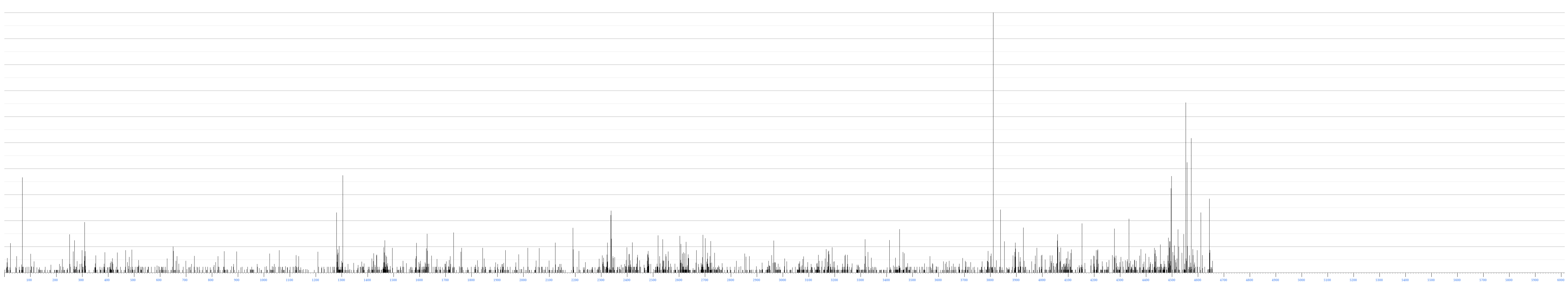
Come si può notare graficamente, gli ultimi due Libri schiacciano in modo catastrofico tutti gli altri, a causa della forza apocalittica del loro segnale. Conviene quindi presentare anche una versione più ristretta con esclusione di questi ultimi, per poter meglio valutare il carattere di tutti gli altri. <u>Il grafico seguente non ha tuttavia oggettività assoluta (poiché, a differenza dello standard, non si auto tara sul massimo assoluto globale della Saga ma ha un'amplificazione artificiosa *ad hoc*) e va inteso solo in senso relativo.</u>

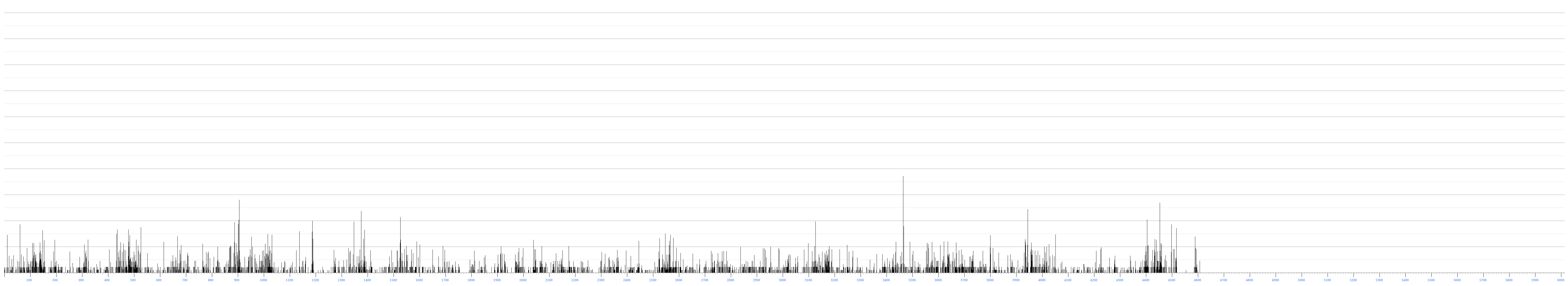


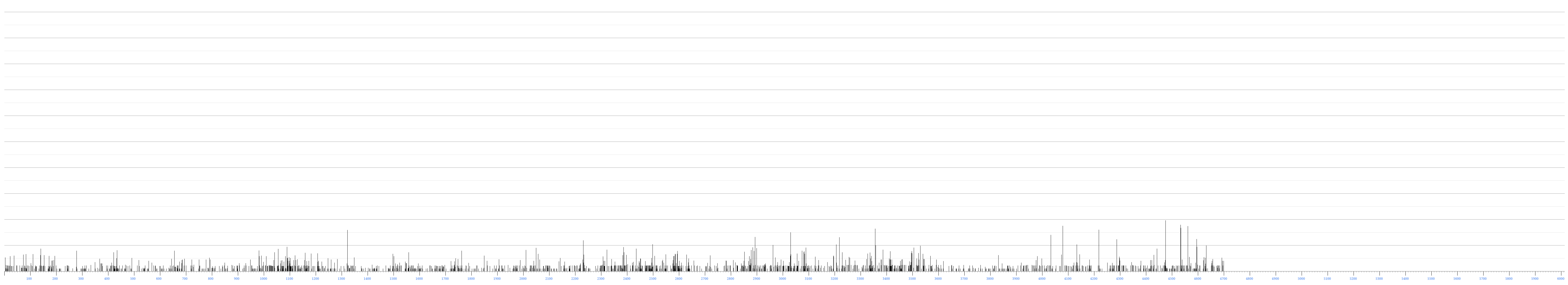












Con questo **diverso fattore di zoom relativo**, emergono alcuni tratti. 1) il rumore di fondo compare, ma rimane relativamente basso, come un’erbetta. Alcuni Libri hanno un segnale debole. I_Venti_del_Giravago ad es. ha solo alcuni slanci nel finale (dopo il ricongiungimento dell’Arca Slavonica con la Fratellanza di origine e la riscoperta delle comuni radici Ortodosse) ed è confrontabile col Il_Contrattacco_Interrogativo (che manca persino di tali slanci : evidentemente il cluster di ricerca non era sufficientemente disgiunto ed intercettava alcune caratteristiche Interrogative). Rocca_di_Stella ha un segnale medio debole ma molto diffuso, segno che da un lato la religione, apocrifa, del Sire Ardente / Sire Fiameggiante è stata sostanzialmente smorzata con successo con la semplice esclusione di questi epiteti dal cluster semantico di ricerca, e tuttavia il segnale persiste ben distinto dal rumore di fondo. La stessa operazione su Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani : l’attribuzione al termine Altissimo del peso minimo, per ridurre la forza dei falsi positivi, dovuti all’esteso uso come aggettivo non sostantivato, ma l’inclusione di Bròndor all’apice delle priorità, ha anche qui smorzato senza cancellare il segnale.
Ne La_Terza_Guerra_Interrogativa il segnale è ancora più debole. Il leitmotiv Interrogativo e del WEIRD / eccentrico / strano / bizzarro lì è talmente potente e tale da escludere quasi ogni altro genere e da non lasciare spazio ad una vera religiosità. G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro ha miriadi di micro clusters, con alcune regioni di macro clusters multipicco intensi.

Il GENERE VAMPIRICO

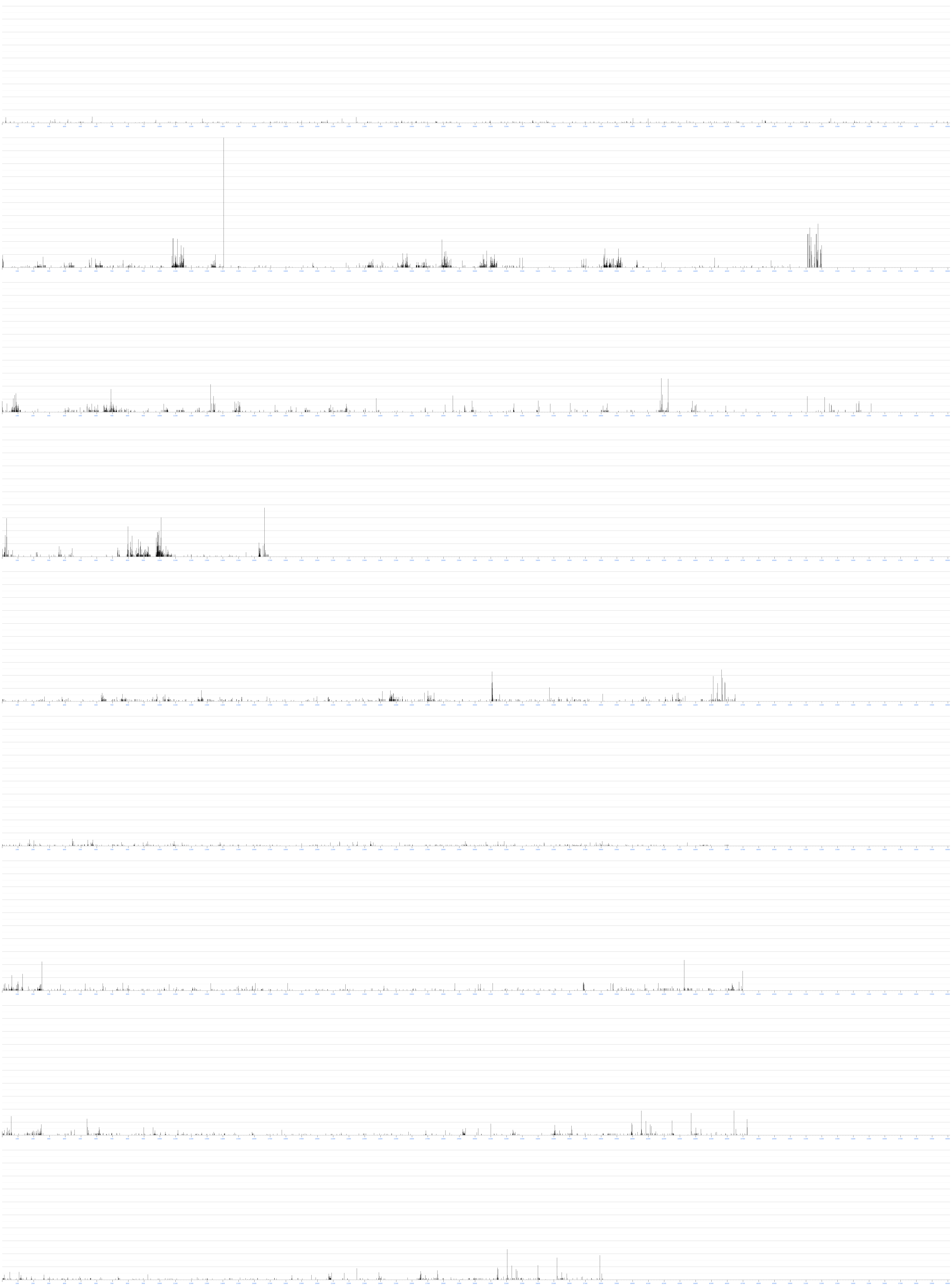
Il Sequel Il_Contrattacco_Interrogativo è stato trattato assai male nella preselezione. Non c'è ancora un glossario stabile, sicché tutti i nomi propri di Vampiroidi, Protovampiri, Vulgroùwl e Zìvouggurd non sono stati inclusi. Questo Libro apparirà quindi gravemente sottopesato (gravemente perché è uno dei due libri dove il Genere Vampiresco dovrebbe essere più intenso, con la tematica delle Ali Nere, che è tanto intensa nel Prequel). Questo cluster di ricerca ha una parte estremamente specifica e praticamente nessuna parte generica *dual–use* e, comunque, è vasto appena la metà (151 lemmi) rispetto al già ristretto cluster GATTOSO / FELINO. Il suo segnale è stato quindi setacciato con una rete molto piccola e risulta proporzionalmente indebolito. Il problema di questo genere nasce da caratteristiche intrinseche non rilevabili dall'algoritmo che si basa su singoli vocaboli. Possiede la nozione della finestra mobile di contesto, ma non è affatto la stessa cosa di offrire supporto nativo flessibile a locuzioni specifiche. Ad esempio sono stati inclusi due aggettivi (*cavo, cave*) ed i sostantivi (*zanne, denti*), con pesi molto bassi per la genericità. Ora isolatamente, magari anche vicini, ma non in funzione sintattica strettamente correlata, queste coppie pesano poco nel punteggio. Eppure la locuzione *zanne cave, canini acuminati*, sono frequentemente associati al genere VAMPIRICO. Anche la parola *sangue* ha ricevuto un peso molto basso perché è assolutamente generica, e il verbo *succhiare* non è stato aggiunto proprio. Tuttavia è chiaro che accostandoli con correlazione sintattica forte (azione *succhiare* complemento oggetto *sangue*) sarebbero una locuzione estremamente diagnostica di episodio di vampirismo, ben oltre a quanto possa premiare l'algoritmo. *Ali* isolatamente non è stato inserito, *membranose e pergamenacee* sì. Ma insieme, *ali membranose*, sono molto più diagnostiche. In sostanza questo genere è largamente sfuggito agli algoritmi, per i loro limiti e per come il genere stesso è stato concretamente implementato nel testo. Non è stato possibile circoscriverlo adeguatamente e ne risulta sottopesato. Di quanto ? Ci è impossibile dirlo, per la stessa ragione per cui non è stato possibile fare compensazioni.

Legenda delle Branche, incluse Bozze e SEQUEL	Cluster VAMPIRICO
1) Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani	6014 elem. Somma Parziale : 180,8467; Max di Riga = 2,8222 - media : 180,8467 / 6014 = 0,0300709511141
2) La_Terza_Guerra_Interrogativa	5200 elem. Somma Parziale : 1418,6872; Max di Riga = 59,8527 - media : 1418,6872 / 5200 = 0,272824461538
3) *Il_Contrattacco_Interrogativo*	5516 elem. Somma Parziale : 811,4872; Max di Riga = 15,6405 - media : 811,4872 / 5516 = 0,14711515591
4) *SEQUEL__G_E_G_di_Napoli*	1694 elem. Somma Parziale : 637,9268; Max di Riga = 22,612 - media : 637,9268 / 1694 = 0,376580165289
5) G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro	4658 elem. Somma Parziale : 685,553; Max di Riga = 14,6559 - media : 685,553 / 4658 = 0,14717754401
6) Rocca_di_Stella	4608 elem. Somma Parziale : 253,2937; Max di Riga = 3,3044 - media : 253,2937 / 4608 = 0,0549682508681
7) I_Venti_del_Giravago	4701 elem. Somma Parziale : 482,6478; Max di Riga = 14,0907 - media : 482,6478 / 4701 = 0,102669176771
8) *SEQUEL__La_Parola_Sepolta*	4683 elem. Somma Parziale : 490,1983; Max di Riga = 11,3447 - media : 490,1983 / 4683 = 0,104676126415
9) La_Parola_Sepolta	3833 elem. Somma Parziale : 342,6501; Max di Riga = 14,077 - media : 342,6501 / 3833 = 0,0893947560657
	Totali : 40957 elementi. **Somma Totale : 5303,2908 - media SAGA :** 5303,2908 / 40957 = **0,129484356764**
	Max assoluto : 59,8527
	Range relativo delle medie : (0,376580165289 – *0,0300709511141) / 0,129484356764 =* ***267,6 %***

Anche questo genere, a fronte della forza modesta (misurabile) del segnale, è tra i più eterogeneamente distribuiti tra i generi rilevati !

Il genere è praticamente assente sia in Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani che in Rocca_di_Stella, privo persino di rumore di fondo e falsi positivi. Rimane molto debole anche in : La_Parola_Sepolta e SEQUEL__La_Parola_Sepolta (laddove c'è un unico Vampiro con un'unica vittima, e solo occasionalmente), I_Venti_del_Giravago (dove ci sono solo pseudovampiri facoltativi, i Mohrlar). Ancora debole, vicino alla media della Saga (il più basso tra i generi misurati), in : Il_Contrattacco_Interrogativo e G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro. Raggiunge un livello sostanzialmente medio forte solo in La_Terza_Guerra_Interrogativa per infine diventare forte in SEQUEL__G_E_G_di_Napoli (le cui bozze sono ancora molto limitate, ma l'avvento dell'Æquòreum e di un'aristocrazia vampira come casta dominante, in seguito al declino degli IronMaster, lascia presupporre che il trendi sia stabile). Per contro i Vampiri del PREQUEL sono piuttosto modesti.

Ecco i dati nel grafico differenziale (non annotato, ma i libri sono sempre nello stesso ordine di tutte le altre figure) :



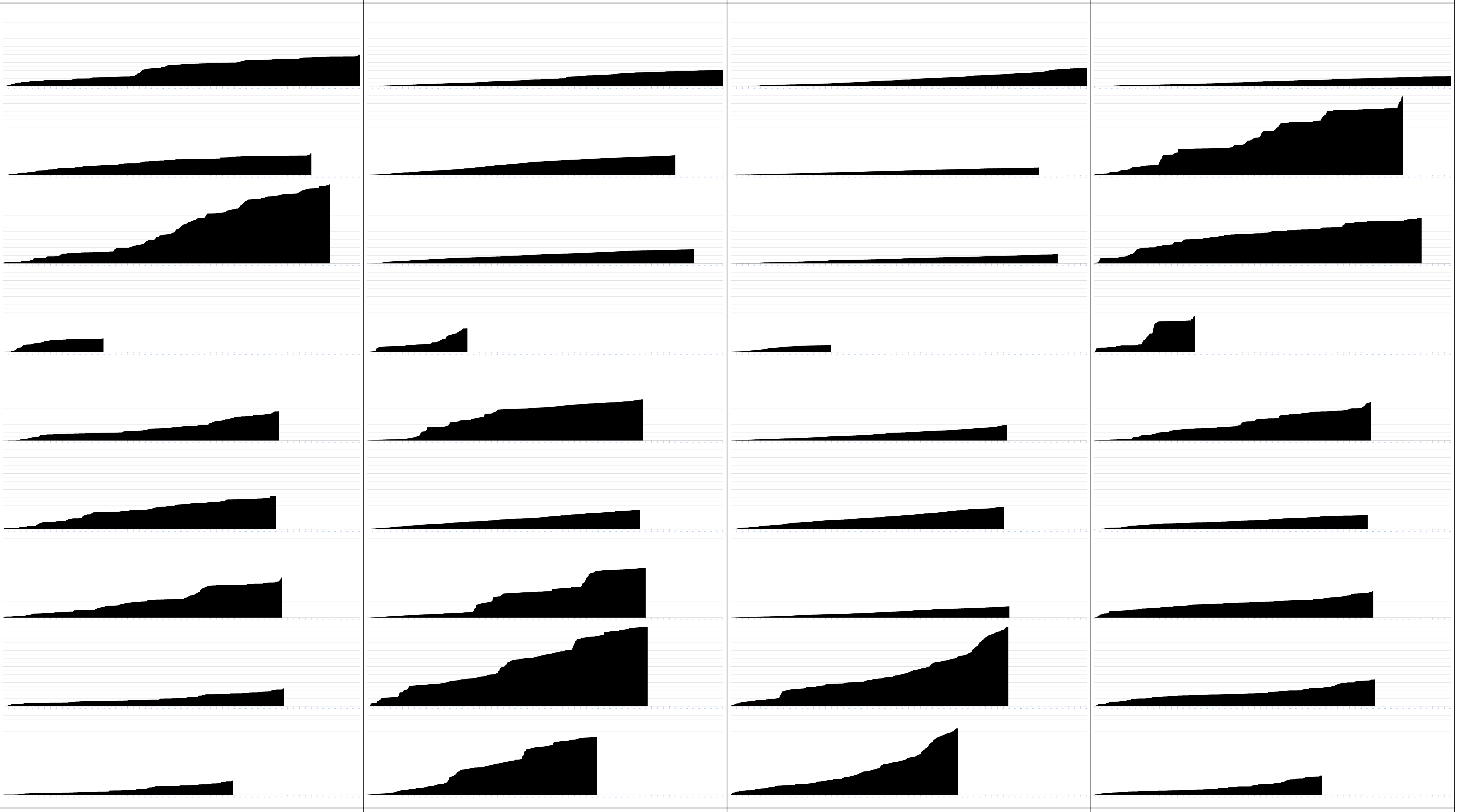
IL GENERE WEIRD (STRANO / ECCENTRICO / BIZZARRO / MISTERO, con tracce “*GOTICHE*”)

Pur essendo presente con estrema intensità nella Saga (anche se non ubiquitario : Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani e Rocca_di_Stella ne contengono molto meno della media), è assolutamente intrattabile dal punto di vista della “semantizzazione” in clusters di ricerca fatti di vocaboli singoli. La stranezza è per sua natura ineffabile, non si origina dai singoli termini ma nasce dalle situazioni assurde ed irrazionali, eventi imprevisti, senza una ragione apparente (si pensi al viaggio di Varel Kajinark e Lady Grayther nei bazar di Hypnobòrea, allo steampunk degli strambi vascelli dei Setacciatori che recuperano relitti misteriosi alla deriva verso il vuoto assoluto oltre l’Arca, alla Natura Interrogativa stessa ed alle entità quadridimensionali, ai misteri dell’Æthéreum, coi suoi Whòldragg, ectoplasmi e fantasmi spettrali), nasce dalle premonizioni superstiziose che aggricciano la pelle della schiena, da un clima arcano, oscuro ed indefinito che si crea accostando molte parole suggestive in un certo ordine, ciascuna presa singolarmente essendo del tutto inadatta a cogliere l’eccentricità. Questo significa che, malgrado ci tenessimo molto, forse persino più di tutti gli altri generi letterari, il SW sviluppato si rivela inadeguato a coglierlo nella sua fuggevolezza. Abbiamo dovuto arrenderci.

I GRAFICI INTEGRALI – comparazioni di aree a colpo d’occhio

Questi grafici rappresentano la sommatoria incrementale degli *scores* di tutti i *text chunks* accumulati sino ad un dato *chunk*. Matematicamente sono “l’integrale” dei grafichi coi picchi precedenti, i quali, naturalmente, a loro volta sono i differenziali di questi ultimi. Siccome la crescita è spesso molto dolce e sfumata e le gobbe si notano poco, la compressione dell’asse longitudinale non solo non guasta ma enfatizza le discontinuità (anche se non in modo così chiaro come nei differenziali), e consente una visione d’insieme della media e dei totali, che danno il timbro totale al libro. La forza totale del segnale è, chiaramente, non l’area totale di ciascuna riga ma solo l’altezza finale. L’area in effetti aumenta non solo a causa dell’elevato valore finale, ma quando gli eventi salienti si verificano in anticipo piuttosto che molto tardi nel testo (e ciò non ha alcun senso). La forza complessiva di un dato genere è la somma delle ordinate finali di ciascun libro, nel complesso della Saga e dell’intera colonna. I grafici non aumentano a velocità uniforme, hanno regioni pianeggianti e poi, a volte, hanno improvvisi sbalzi : quelli sono gli eventi, talvolta i capitoli, dove succede qualcosa di particolarmente caratterizzante per il dato genere misurato.
I grafici integrali dei tra generi a confronto (i Libri si “leggono” in orizzontale) :

GATTOSO / FELINO	EROTICO / PORNOGRAFICO	RELIGIOSO / SPIRITUALE / MISTICO / ULTRATERRENO / METAFISICO	VAMPIRICO



Score : 16765,214 **(306 lemmi cercati)** **massimo assoluto : 53,257**	**Score : 19283,1132** **(893 lemmi cercati)** **massimo assoluto : 65,9022**	**Score : 31332,9443** **(896 lemmi cercati)** **massimo assoluto : 159,0133**	**Score : 5303,2908** **(151 lemmi cercati)** **massimo assoluto : 59,8527**

Si consideri attentamente il fatto che questi insiemi di grafici sono rinormalizzati internamente ed esternamente in modo da avere la stessa scala ASSOLUTA (al netto però, come già riferito, della diversa numerosità del Cluster Semantico che ha setacciato il testo : il secondo e terzo cluster sono di circa 900 termini l’uno, anche se il secondo ha più termini dual–use e quindi tende a raccogliere più rumore bianco, ma il primo è di appena 300 termini (tra l’altro più specifici) che tendono di per sé a sottopesare e l’ultimo solo 150). Quindi i grafici oltre che in verticale possono essere confrontati in orizzontale, tra colonne diverse. Ciò significa che si possono confrontare visivamente gli *exit points* entro ciascuna colonna, ma non tra colonne diverse : il fattore amplificante è diverso ! La quarta colonna, che non sembrerebbe poi tanto più debole della seconda, il realtà lo è eccome. La terza colonna, che in figura sembra la meno carica di nero, in valore assoluto vale uno score totale di oltre 31000, ben sei volte l’ultima colonna (5300). A metà strada stanno, in ordine crescente, la colonna 1 e la 2.
Per avere una percezione corretta (ma sempre al netto della diversa dimensione del *set* di termini di ricerca, differenza non compensabile a posteriori in alcun modo), del fattore di zoom reale, l’altezza relativa REALE dei grafici si potrebbe ottenere lasciando invariata la terza colonna (genere RELIGIOSO / SPIRITUALE), e riducendo l’altezza della prima colonna di un fattore 16765,214 / 31332,9443 = 0,5351 (poco più della metà), riducendo l’altezza della seconda colonna di un fattore 19283,1132 / 31332,9443 = 0,6154 (il 61,5 % della figura) e la quarta colonna dovrebbe essere ridotta in altezza di un fattore 5303,2908 / 31332,9443 = 0,16926 (appena il 16,3 % della figura o poco più di un sesto dell’altezza mostrata).

I due libri più equi (qualunque cosa ciò significhi) appaiono Rocca_di_Stella e Cronache_di_Constance_e_delle_Stirpi_Originarie_dei_Guardiani, il più sbilanciato in assoluto è Il_Contrattacco_Interrogativo, con una singola componente intensa, mentre le due La_Parola_Sepolta, I_Venti_del_Giravago e G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro hanno due componenti intense ed una debole (rispettivamente quella GATTOSA e quella RELIGIOSA). G_E_G_di_Napoli__Gaetano_Esposito_de_Gennaro ed I_Venti_del_Giravago sono quantitativamente piuttosto simili anche se mostrano interalmente sequenze temporali differenti.

Conclusione :

Per l’agognato e famigerato GENERE WEIRD (STRANO / ECCENTRICO / BIZZARRO / MISTERO, con tracce “*GOTICHE*”) … beh, occorrerà aspettare recensioni altrui.
Per chi volesse acquistare gli Ebooks dei 6 Libri della Saga (esclusi i tre Sequel), sono disponibili in vari Estore online :
https://sgrpb7.gumroad.com/
https://payhip.com/PluriLogiaBySGRPB7
https://plurilogiaunafamigliadilibriparalleli.sell.app/
https://www.amazon.com/dp/B0GNTCJ5X2?binding=kindle_edition&ref=dbs_dp_sirpi
https://www.amazon.com/stores/SGRPB7-Studio-Giochi-di-Ruolo-Poly-Books-7

by Studio_Giochi_di_Ruolo_Poly_Books_7
(revisori di bozze : Pàncer “Doréamon”, Dæmònya, Adenore Castrovizzardo)